Source-Free Domain Adaptation via Distribution Estimation

Ning Ding[1], Yixing Xu[2], Yehui Tang[1,2], Chao Xu[1], Yunhe Wang[2]*, Dacheng Tao[3]
[1] Key Lab of Machine Perception (MOE), School of Artificial Intelligence, Peking University
[2] Huawei Noah's Ark Lab [3] JD Explore Academy, China
dingning@stu.pku.edu.cn, {yixing.xu, yunhe.wang}@huawei.com, yhtang@pku.edu.cn
xuchao@cis.pku.edu.cn, dacheng.tao@gmail.com

Abstract

*Domain Adaptation aims to transfer the knowledge learned from a labeled source domain to an unlabeled target domain whose data distributions are different. However, the training data in source domain required by most of the existing methods is usually unavailable in real-world applications due to privacy preserving policies. Recently, Source-Free Domain Adaptation (SFDA) has drawn much attention, which tries to tackle domain adaptation problem without using source data. In this work, we propose a novel framework called SFDA-DE to address SFDA task via source **D**istribution **E**stimation. Firstly, we produce robust pseudo-labels for target data with spherical k-means clustering, whose initial class centers are the weight vectors (anchors) learned by the classifier of pretrained model. Furthermore, we propose to estimate the class-conditioned feature distribution of source domain by exploiting target data and corresponding anchors. Finally, we sample surrogate features from the estimated distribution, which are then utilized to align two domains by minimizing a contrastive adaptation loss function. Extensive experiments show that the proposed method achieves state-of-the-art performance on multiple DA benchmarks, and even outperforms traditional DA methods which require plenty of source data.*

1. Introduction

In the past few years, deep Convolutional Neural Networks (CNNs) have achieved remarkable performance on many visual tasks such as classification [19], object detection [8], semantic segmentation [28], *etc.* However, the success of CNNs relies heavily on the hypothesis that the distributions of the training data is identical to that of the test data. Thus, models trained with data from a certain scenario (source domain) can hardly generalize well to other real-world application scenarios (target domains), and may suffer from severe performance drop. Moreover, the difficulty of collecting enough labeled training data also hinders CNNs from directly learning with target domain data. Unfortunately, CNNs deployed in real-world scenarios always encounter new situations, such as the change of weather and variation of illumination in autonomous driving.

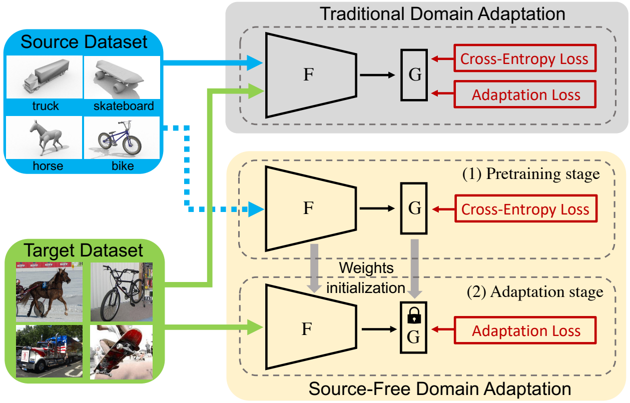


Figure 1. (Top) Traditional domain adaptation methods require data from both source domain and target domain simultaneously. (Bottom) In source-free domain adaptation, source data can only be used in the pretraining stage, and cannot be accessed in the later adaptation stage. Adaptation is achieved by utilizing target data and the model pretrained on source domain.

Therefore, a lot of attention is paid to the domain shift problem [1, 43] mentioned above, and *Domain Adaptation* (DA) theory has been developed to solve it. DA algorithms directly help deep models transfer knowledge learned from a fully annotated source domain to a separately distributed target domain whose annotations are entirely unavailable. Existing advances in deep learning-based DA methods [4, 24, 30, 34] generally achieve model transferability by means of mapping two different data distributions simultaneously into a mutual feature space shared across two domains.

However, people are getting more aware of the importance of privacy data protection nowadays. Strict policies regarding data privacy concerns have been published all

* Corresponding author.

around the world. More AI companies also choose to open source their pretrained models only, yet keep the source dataset used for training unreleased [46]. Therefore, most of the traditional DA methods become infeasible to transfer knowledge to target domain when source data is no longer accessible since these methods basically assume that data from both source domain and target domain is available.

To overcome this data-absent problem, some recent works [18,20,23,25,57] explored more general approaches to achieve domain adaptation without accessing source data. Only unlabeled target domain data and the model pretrained on source domain are required to accomplish the cross-domain knowledge transfer. Such a new unsupervised learning setting for domain adaptation task is called *Source-Free Domain Adaptation* (SFDA). SHOT [25] utilizes information maximization and entropy minimization. 3C-GAN [23] uses a generative model to enrich target data to enhances model performance. G-SFDA [58] learns different feature activations by exploiting neighborhood structure of target data. A^{2}Net [52] introduces a new classifier and adopt adversarial training to align two domains. Despite the fact that these SFDA methods utilize the source domain knowledge contained by the pretrained model, none of them explicitly align the distributions between source domain and target domain to achieve adaptation.

In this paper, we focus on image classification task under SFDA setting. We manage to estimate the source distribution without accessing source data. Specifically, we utilize the domain information captured by the model pretrained on source data and treat the weights learned by source classifier as class anchors. Then, these anchors are used as the initialization of feature center for each class and spherical k-means is performed to cluster target features in order to produce robust pseudo-labels for target data. Furthermore, we dynamically estimate the feature distributions of source domain class-wisely by utilizing the semantic statistics of target data along with their corresponding anchors, which is called Source Distributions Estimation (**SDE**). Finally, we sample surrogate features from distributions derived from SDE to simulate the real but unknown source features, and then align them with target features by minimizing a contrastive adaptation loss function to facilitate source-free domain adaptation. In short, if the feature distribution of target domain is well-aligned with source domain, the source classifier will naturally adapt to the target domain data.

We validate our proposed SFDA-DE method on three public DA benchmarks: Office-31 [38], Office-Home [50] and VisDA-2017 [37]. Experiment results show that the proposed SFDA-DE method achieves state-of-the-art performance on Office-Home (72.9%) and VisDA-2017 (86.5%) among all SFDA methods, and is even superior to some recently proposed traditional DA methods that require accessing source domain data.

2. Related Work

Traditional domain adaptation. Domain Adaptation (DA) as a research topic has been studied for a long time [1]. With the emergence of deep learning [19, 41], CNNs with superior capacity to capture high level features become the first choice to perform adaptation. As a result, many related tasks have been developed in the field of visual DA, such as multi-source DA [36, 53], semi-supervised DA [11, 39], partial DA [2], open set DA [27, 35], universal DA [40], *etc*. DA aims to improve the generalizability of a model which is learned on a labeled source domain. When fed with data drawn from a different target distribution, model performance declines drastically. This is referred to as *covariate shift* [42–44] or *domain shift* problem. To tackle this problem, lots of methods try to align feature distributions of different domains via minimizing Maximum Mean Discrepancy (MMD) [29,31,32,43], which is a non-parametric kernel function embedded into reproducing kernel Hilbert space (RKHS) to measure the difference between two probability distributions [9, 15]. Moreover, Kang *et al.* [17] incorporates contrastive learning technique [3] into MMD-based method to further boost model transferability. Meanwhile, Zellinger *et al.* [60] and Sun *et al.* [45] propose to align high order statistics captured by networks like central moment to achieve domain adaptation. Apart from directly aligning two distributions, some recent works [7,30,49] employ adversarial training by adding an extra feature discriminator. In this way, the networks are forced to learn domain-invariant features to confuse the discriminator.

Source-free domain adaptation. All the methods mentioned above expect both labeled source data and unlabeled target data to achieve domain adaptation process, which is often impractical in real-world scenario. In most cases, one can only access the unlabeled target data and the model pretrained by source data. To this end, some recent works [20, 23, 25, 26, 48, 52, 56, 58] regarding source-free domain adaptation emerge. These methods provide solutions to adapt the model to unseen domains without using original training data. SHOT [25] utilizes information maximization and entropy minimization via pseudo-labeling strategy to adapt the trained classifier to target features. [23, 26] both use generative models to model the distribution of target data by generating target-style images to enhance the model performance on target domain. G-SFDA [58] forces the network to activate different channels for different domains while paying attention to the neighborhood structure of data. A^{2}Net [52] introduces a new target classifier to align two domains via adversarial training manner. SoFA [59] uses a Variational Auto-Encoder to encode target distribution in latent space while reconstructing the target data in image space to constrain the latent features. Many of the above methods freeze the source classifier during adaptation to preserve class information, and

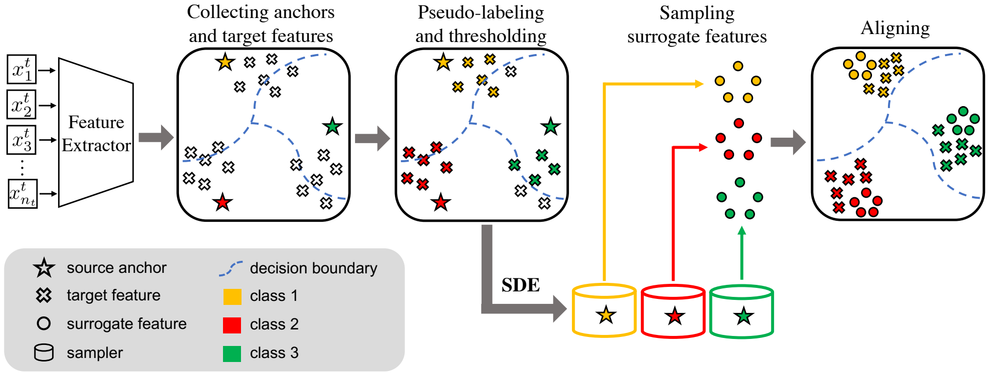


Figure 2. Overall training pipeline of our proposed SFDA-DE method.

assign pseudo-labels based on the classifier's output. Here we follow the idea of freezing the source classifier [25] but use a more robust pseudo-labeling strategy via spherical k-means clustering. Moreover, we propose Source Distribution Estimation (SDE), aiming to approximate the source feature distribution without accessing the source data. After that, the target distribution can be directly aligned with the estimated distribution to adapt to the source classifier.

3. Method

In this section, we first describe the problem setting for source-free domain adaptation and notations to be used afterward. Then we elaborate our proposed SFDA-DE method in three steps to address SFDA problem. First of all, we obtain robust pseudo-labels for target data by utilizing source anchors and spherical k-means clustering. Secondly, we estimate the class-conditioned feature distribution of source domain. Finally, surrogate features are sampled from the estimated distribution to align two domains by minimizing a contrastive adaptation loss function.

3.1. Preliminaries and notations

In this paper, we use $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ to denote the source domain dataset with n_s labeled samples, where $y \in \mathcal{Y} \subseteq \mathbb{R}^K$ is the one-hot ground-truth label and K is the total number of classes of the label set $\mathcal{C} = \{1, 2, \cdots, K\}$. $\mathcal{D}_t = \{(x_i^t)\}_{i=1}^{n_t}$ denotes the target domain dataset with n_t unlabeled samples which has the same underlying label set $\mathcal{C}$ as that of $\mathcal{D}_s$. In SFDA scenario, we have access to the model $\mathbf{G}(\mathbf{F}(\cdot))$ which is already pretrained on $\mathcal{D}_s$ in a supervised manner by cross-entropy loss, where $\mathbf{F}$ denotes the CNN feature extractor followed by a linear classifier $\mathbf{G}$. During training, only data in $\mathcal{D}_t$ is available and no data in $\mathcal{D}_s$ can be used. Besides, we use $f = \mathbf{F}(x) \in \mathbb{R}^m$ to denote m-dimensional feature representations and use $\mathbf{w}^G \in \mathbb{R}^{m \times K}$ to denote the weights learned by $\mathbf{G}$, where $\mathbf{w}_k^G \in \mathbb{R}^m$ is the k-th weight vector of $\mathbf{w}^G$.

3.2. Pseudo-labeling by exploiting anchors

In many works, pseudo-labeling is an important technique to obtain category information for those unlabeled samples and is usually realized by exploiting the highly-confident outputs derived by the classifier. However in SFDA task, the classifier $\mathbf{G}$ is pretrained on source domain data and will encounter the distribution shift problem when classifying target domain data. Therefore, it's crucial to find a robust way to solve the distribution shift problem and assign correct labels to unlabeled target data. Thus, we consider obtaining pseudo-labeling via spherical k-means.

Specifically, given a class label predicted by the linear classifier $\mathbf{G}$ as

$$\hat{y}_i = \arg\max_k f_i^\top \mathbf{w}_k^G \ , \ k \in \mathcal{C} = \{1, 2, \cdots, K\} \ , \qquad (1)$$

where $\hat{y}_i \in \mathbb{R}^K$ is the logits vector before softmax. Note that each element in the class probability vector is derived by the dot product between the feature and each weight vector of the classifier. Thus, data of the k-th class tends to yield feature representation that activates the k-th weight vector in $\mathbf{G}$. Features of data from the k-th class should gather around $\mathbf{w}_k^G$. Therefore, $\mathbf{w}_k^G$ can be treated as an **anchor** of the k-th class which contains overall characteristics that represent the whole k-th class.

In SFDA task, target features would drift away from source anchors which makes it hard to directly predict labels for target data with pretrained classifier $\mathbf{G}$. Thus, we propose to assign pseudo-labels for target data via spherical k-means. We first cluster the target data by setting anchors as the initial cluster centers: $\mathcal{A}_k^{(0)} = \mathbf{w}_k^G$. Then we perform spherical k-means iteratively between (1) assigning pseudo-labels via minimum-distance classifier: $\hat{y}_i^t = \arg\min_k Dist(\mathcal{A}_k^{(m)}, f_i^t)$ and (2) computing new cluster centers $\mathcal{A}_k^{(m+1)} = \frac{\sum_{i=1}^{n_t} \mathbb{1}(\hat{y}_i^t = k) f_i^t}{\sum_{i=1}^{n_t} \mathbb{1}(\hat{y}_i^t = k)}$, where $Dist(\mathbf{a}, \mathbf{b}) = \frac{1}{2}(1 - \frac{\mathbf{a}^\top \mathbf{b}}{|\mathbf{a}| \cdot |\mathbf{b}|})$ is the cosine distance, m denotes the number of current iterations and $\mathbb{1}(\cdot)$ is the indicator function. Iteration will stop when all class centers converge. After clustering is done, a confidence threshold $\tau \in (0, 1)$ is set to filter out ambiguous samples so that a confidently pseudo-labeled target dataset $\mathcal{D}'_t$ is constructed:

$$\mathcal{D}'_t = \{(x_i^t, \hat{y}_i^t) \mid Dist(f_i^t, \mathcal{A}_{\hat{y}_i^t}) < \tau, \ \hat{y}_i^t \in \mathcal{C}\}_{i=1}^{n'_t}. \qquad (2)$$

Given the robust pseudo-labels derived above, we use $x_{i,k}^t$ to denote the target data x_i^t with pseudo-label $\hat{y}_i^t = k$, and use $f_{i,k}^t = \mathbf{F}(x_{i,k}^t)$ to denote its corresponding feature representation in the following of this paper. Similar to the idea proposed in [25], we freeze $\mathbf{G}$ to fix the source anchors in order to stabilize the adaptation to target domain.

3.3. Source Distribution Estimation

In traditional DA setting, feature distributions of data from both source and target domains can be estimated by mini-batch sampling from $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively. Then the target distribution can be explicitly aligned with the source one and classified by the pretrained source classifier $\mathbf{G}$ [17, 31, 32]. However, source data is unavailable in SFDA setting, which makes it impossible to know the source distribution. To tackle this problem, Yang *et al.* [58] focuses on neighborhood structure and channel activation. Liang *et al.* [25] exploits information maximization and self-supervision to implicitly align feature representations. Nevertheless, none of the existing methods address SFDA problem by explicitly aligning the source distribution with the target distribution, and thus achieve sub-optimal results. We manage to explicitly estimate the source feature distribution without accessing source data by presenting Source Distribution Estimation (SDE) method.

Concretely, we assume feature representations of source domain follow a class-conditioned multivariate Gaussian distribution $f_{i,k}^s \sim \mathcal{N}_k^s(\mu_k^s, \Sigma_k^s)$, where $f_{i,k}^s = \mathbf{F}(x_i^s|y_i^s{=}k)$ and $k{\in}\mathcal{C}{=}\{1, 2, \cdots, K\}$. Essentially, μ_k^s can be viewed as the center of feature representations of the k-th class data in source domain and Σ_k^s is the covariance matrix which captures the variation in features of the k-th class and contains rich semantic information [51]. Then we can use a surrogate distribution $\mathcal{N}_k^{sur}(\hat{\mu}_k^s, \hat{\Sigma}_k^s)$ to approximate the actual but unknown source distribution $\mathcal{N}_k^s$ for each class $k \in \mathcal{C}$.

A good estimator for μ_k^s should be discriminative enough and reflect the intrinsic characteristics of the k-th class data in source domain. If we directly use the feature mean of target data of the k-th class $\bar{f}_k^t = \frac{\sum_i f_{i,k}^t}{\sum_{x_i^t \in \mathcal{D}_t'} \mathbb{1}(\hat{y}_i^t = k)}$ as an estimator for μ_k^s, obviously the above conditions cannot be satisfied due to the existence of domain shift problem. Recall the observation in Sec. 3.2 that anchors contain overall characteristics of the corresponding class. Thus, we propose to utilize anchors to calibrate the estimator for mean of the surrogate source distribution:

$$\hat{\mu}_k^s = \|\bar{f}_k^t\|_2 \cdot \frac{\mathbf{w}_k^G}{\|\mathbf{w}_k^G\|_2}\ ,\ k \in \mathcal{C}, \tag{3}$$

which implies that the direction of estimated source feature mean is the same as the corresponding anchor but the scale of it is derived from target features. Another reason for the calibration is that there is usually a difference in norm between anchors and features, which is $\|\mathbf{w}_k^G\|_2 < \|f_{i,k}^t\|_2 \approx \|f_{i,k}^s\|_2$, empirically. Therefore, it's not appropriate to directly use anchors as the estimator of mean either.

As for covariance matrices, many works [5, 22, 24, 51] study the statistics of deep features and reveal that class-conditioned covariance implies the activated semantic directions and correlations between different feature channels. We assume that the intra-class semantic information of target features is roughly consistent with that of the source. Hence we derive the estimator for source covariance Σ_k^s from statistics of target features:

$$\hat{\Sigma}_k^s = \gamma \cdot \Sigma_k^t = \gamma \cdot \frac{\mathbf{f}_k^t \cdot {\mathbf{f}_k^t}^\top}{\sum\limits_{x_i^t \in \mathcal{D}_t'} \mathbb{1}(\hat{y}_i^t = k)}, \tag{4}$$

where $\mathbf{f}_k^t = [f_{1,k}^t - \bar{f}_k^t, \cdots, f_{i,k}^t - \bar{f}_k^t, \cdots]$ is a matrix whose columns are centralized target features of the k-th class in $\mathcal{D}_t'$. We use a controlling coefficient γ to adjust the sampling range and semantic diversity of sampled surrogate features. Details of selecting γ will be studied in Sec. 4.3.

By exploiting anchors and target features, we derive K class-conditioned surrogate source distributions

$$\mathcal{N}_k^{sur}(\|\bar{f}_k^t\|_2 \frac{\mathbf{w}_k^G}{\|\mathbf{w}_k^G\|_2}, \frac{\gamma \cdot \mathbf{f}_k^t \cdot {\mathbf{f}_k^t}^\top}{\sum\limits_{x_i^t \in \mathcal{D}_t'} \mathbb{1}(\hat{y}_i^t = k)}), k \in \mathcal{C}, \tag{5}$$

from which we can sample surrogate features $f_k^{sur} \sim \mathcal{N}_k^{sur}(\hat{\mu}_k^s, \hat{\Sigma}_k^s)$ to simulate the real source features.

3.4. Source-free domain adaptation

In the previous section, we are able to estimate the source distribution without accessing source data by exploiting domain knowledge preserved in the pretrained model with the proposed SDE method. Thus, we can sample data from the estimated distribution as surrogate source data, and the SFDA problem becomes the traditional DA problem. We adopt Contrastive Domain Discrepancy (CDD) introduced by Kang *et al.* [17] to explicitly align the target distribution with the estimated source distribution.

Specifically, we choose a random subset $\mathcal{C}' \subset \mathcal{C}$ from the label set $\mathcal{C}{=}\{1, 2, \cdots, K\}$ before each forward pass. Then for each $k \in \mathcal{C}'$, we sample n_b target images from $\mathcal{D}_t'$ to construct a set of data $\{\{(x_i^t, \hat{y}_i^t = k)\}_{i=1}^{n_b}|k \in \mathcal{C}'\}$ and derive the target mini-batch $\{\{f_{i,k}^t = \mathbf{F}(x_i^t|\hat{y}_i^t{=}k)\}_{i=1}^{n_b}|k \in \mathcal{C}'\}$. Correspondingly, we sample n_b features from surrogate source distributions for each $k \in \mathcal{C}'$ to construct the source mini-batch $\{\{f_{j,k}^{sur} \sim \mathcal{N}_k^{sur}\}_{j=1}^{n_b}|k \in \mathcal{C}'\}$. Therefore, for any two class $k_1, k_2 \in \mathcal{C}'$, a class-conditioned version of MMD that measures discrepancy between surrogate source distribution and target distribution is defined as

$$\begin{aligned}\mathcal{L}_{\text{MMD}}^{k_1,k_2} = &\sum_{i=1}^{n_b}\sum_{j=1}^{n_b} \frac{\Bbbk(f_{i,k_1}^{sur}, f_{j,k_1}^{sur})}{n_b \cdot n_b} + \sum_{i=1}^{n_b}\sum_{j=1}^{n_b} \frac{\Bbbk(f_{i,k_2}^{t}, f_{j,k_2}^{t})}{n_b \cdot n_b} \\ &-2\sum_{i=1}^{n_b}\sum_{j=1}^{n_b} \frac{\Bbbk(f_{i,k_1}^{sur}, f_{j,k_2}^{t})}{n_b \cdot n_b},\end{aligned} \tag{6}$$

Algorithm 1: SFDA training process within one epoch

```
Input: unlabeled target images $\{x_i^t\}_{i=1}^{n_t}$, label set $\mathcal{C}$,
       pretrained feature extractor $\mathbf{F}$, frozen
       classifier $\mathbf{G}$, confidence threshold $\tau$, number
       of iterations $t$.
Initialize the cluster center with source anchors $\mathbf{w}_k^G$
  learned by $\mathbf{G}$ for each $k \in \mathcal{C} = \{1, 2, \cdots, K\}$;
Apply spherical k-means on target features and
  construct confident pseudo-labeled set $\mathcal{D}_t'$ with $\tau$;
Perform SDE to derive K surrogate source
  distributions $\mathcal{N}_k^{sur}(\hat{\mu}_k^s, \hat{\Sigma}_k^s)$ according to Eq. (5);
for $i = 1, 2, \ \ldots, t$ do
    Sample target mini-batch
      $\{\{f_{i,k}^t = \mathbf{F}(x_{i,k}^t)\}_{i=1}^{n_b} | k \in \mathcal{C}'\}$;
    Sample source mini-batch
      $\{\{f_{j,k}^{sur} \sim \mathcal{N}_k^{sur}\}_{j=1}^{n_b} | k \in \mathcal{C}'\}$;
    Compute CDD loss according to Eq. (7);
    Do backward and update weights of $\mathbf{F}$.
end
```

where $\Bbbk(\cdot, \cdot)$ is kernel functions that embeds feature representations in Reproducing Kernel Hilbert Space (RKHS). Utilizing the data in both source batch and target batch, the CDD loss is calculated as

$$\mathcal{L}_{\text{CDD}} = \frac{\sum\limits_{k \in \mathcal{C}'} \mathcal{L}_{\text{MMD}}^{k,k}}{|\mathcal{C}'|} - \frac{\sum\limits_{k_1 \in \mathcal{C}'} \sum\limits_{k_2 \in \mathcal{C}'}^{k_1 \neq k_2} \mathcal{L}_{\text{MMD}}^{k_1,k_2}}{|\mathcal{C}'|(|\mathcal{C}'| - 1)}, \tag{7}$$

in which the first term represents intra-class domain discrepancy to be diminished and the second represents inter-class domain discrepancy to be enlarged. By explicitly treating data from different classes as negative sample pairs, CDD loss facilitates intra-class compactness and inter-class separability, which is beneficial to learning discriminative target features.

Algorithm 1 shows the overall training process of our proposed SFDA method within one epoch. As the adaptation proceeds, target features are driven closer and closer to approach anchors and the statistics of target features will change constantly. Therefore, we perform both pseudo-labeling method in Sec. 3.2 and SDE method in Sec. 3.3 at the beginning of every epoch to dynamically re-estimate the surrogate source distribution $\mathcal{N}_k^{sur}$.

4. Experiments

In this section, we first validate the effectiveness of the proposed SFDA-DE method based on three benchmarks. Then we conduct extensive experiments on hyper-parameter selection, ablation study, visualization, *etc*.

Table 1. Classification accuracy (%) on Office-31 dataset for source-free domain adaptation (ResNet-50). Our method achieves state-of-the-art performance on A→D and A→W tasks. Best results under SFDA setting are shown in bold font.

Method	source -free	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
MDD [61]	×	93.5	94.5	74.6	98.4	72.2	100.0	88.9
GVB-GD [4]	×	95.0	94.8	73.4	98.7	73.7	100	89.3
MCC [16]	×	95.6	95.4	72.6	98.6	73.9	100	89.4
GSDA [13]	×	94.8	95.7	73.5	99.1	74.9	100	89.7
CAN [17]	×	95.0	94.5	78.0	99.1	77.0	99.8	90.6
SRDC [47]	×	95.8	95.7	76.7	99.2	77.1	100	90.8
SFDA [18]	✓	92.2	91.1	71.0	98.2	71.2	99.5	87.2
SHOT [25]	✓	94.0	90.1	74.7	98.4	74.3	99.9	88.6
3C-GAN [23]	✓	92.7	93.7	75.3	98.5	**77.8**	99.8	89.6
A^{2}Net [52]	✓	94.5	94.0	**76.7**	**99.2**	76.1	**100**	**90.1**
SFDA-DE (ours)	✓	**96.0**	**94.2**	76.6	98.5	75.5	99.8	**90.1**

4.1. Experimental settings

Office-31 dataset. Office-31 [38] is a small-scale benchmark with 3 domains, **A**mazon (2,817), **D**SLR (498) and **W**ebcam (795) . There are totally 4,110 images belonging to 31 categories collected from real world scenarios.

Office-Home dataset. Office-Home [50] is a complex benchmark comprised of four visually-dissimilar domains: **Ar**tistic images, **Cl**ipart images, **Pr**oduct images, and **Re**al-world images. This dataset contains 12 transfer tasks and a total number of 15,500 images from 65 classes.

VisDA-2017 dataset. VisDA-2017 [37] is a large-scale synthetic-to-real dataset with 12 classes in both domain. The synthetic domain contains 150K rendered 3D images with various poses and lighting conditions. The corresponding real domain contains about 55K real-world images.

Pretraining on source domain. We use momentum SGD optimizer with exponential decay learning rate schedule $\eta = \eta_0(1 + \alpha \cdot i)^{-\beta}$, where η_0 is the initial learning rate and i is the training steps. Weight decay is set to 5e-4 and momentum is set to 0.9. For Office-31 and Office-Home dataset, we employ ResNet-50 [12] as our feature extractor $\mathbf{F}$ and a single fully-connected layer as classifier $\mathbf{G}$. We set $\eta_0 = 0.001$, $\alpha = 0.001$ and $\beta = 0.75$. The model is trained for 50 epochs on all source domains. For VisDA-2017 dataset, we employ ResNet-101 as the feature extractor $\mathbf{F}$ and train it for 500 steps on source domain. We set $\eta_0 = 0.001$, $\alpha = 0.0005$ and $\beta = 2.25$. For all 3 datasets, the learning rate of $\mathbf{G}$ is set to be 10 times bigger than $\mathbf{F}$ and the batch size is set to 64 for all domains. The source dataset is randomly split into a training set accounting for 90% and a validation set accounting for 10% in order to guarantee the model converges.

SFDA implementation detail. We follow the standard SFDA setups adopted by [25, 52]. We use all weights of the pretrained model as initialization and freeze all anchors $\mathbf{w}_k^G$ in classifier $\mathbf{G}$ during SFDA stage. For Office31 and Office-Home dataset, we use the same optimization setting and learning rate schedule as the aforementioned pretraining stage. We empirically set $\tau = 0.6$, $\gamma = 1$, $|\mathcal{C}'| = 12$

Table 2. Classification accuracy (%) on Office-Home dateset for source-free domain adaptation (ResNet-50). Our method achieves state-of-the-art performance. Best results under SFDA setting are shown in bold font.

Method	source-free	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
GSDA [13]	×	61.3	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [4]	×	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA [10]	×	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
TSA [24]	×	57.6	75.8	80.7	64.3	76.3	75.1	66.7	55.7	81.2	75.7	61.9	83.8	71.2
SRDC [47]	×	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
FixBi [34]	×	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
SFDA [18]	✓	48.4	73.4	76.9	64.3	69.8	71.7	62.7	45.3	76.6	69.8	50.5	79.0	65.7
G-SFDA [58]	✓	57.9	78.6	81.0	66.7	77.2	77.2	65.6	56.0	82.2	72.0	57.8	83.4	71.3
SHOT [25]	✓	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
A^{2}Net [52]	✓	58.4	79.0	**82.4**	67.5	**79.3**	78.9	**68.0**	56.2	**82.9**	**74.1**	60.5	85.0	72.8
SFDA-DE (ours)	✓	**59.7**	**79.5**	**82.4**	**69.7**	78.6	**79.2**	66.1	**57.2**	82.6	73.9	**60.8**	**85.5**	**72.9**

Table 3. Per-class accuracy and mean accuracy (%) on VisDA-2017 dateset for source-free domain adaptation (ResNet-101). Our method achieves state-of-the-art performance. Best results under SFDA setting are shown in bold font.

Method	source-free	plane	bike	bus	car	horse	knife	mcycle	person	plant	sktbrd	train	truck	Avg.
SFAN [54]	×	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
SWD [21]	×	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
MCC [16]	×	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
STAR [33]	×	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
RWOT [55]	×	95.1	80.3	83.7	90.0	92.4	68.0	92.5	82.2	87.9	78.4	90.4	68.2	84.0
SE [6]	×	95.9	87.4	85.2	58.6	96.2	95.7	90.6	80.0	94.8	90.8	88.4	47.9	84.3
SFDA [18]	✓	86.9	81.7	84.6	63.9	93.1	91.4	86.6	71.9	84.5	58.2	74.5	42.7	76.7
3C-GAN [23]	✓	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
SHOT [25]	✓	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
A^{2}Net [52]	✓	94.0	87.8	85.6	66.8	93.7	95.1	85.8	81.2	91.6	88.2	86.5	56.0	84.3
G-SFDA [58]	✓	**96.1**	88.3	**85.5**	**74.1**	**97.1**	95.4	**89.5**	79.4	95.4	92.9	**89.1**	42.6	85.4
SFDA-DE (ours)	✓	95.3	**91.2**	77.5	72.1	95.7	**97.8**	85.5	**86.1**	**95.5**	**93.0**	86.3	**61.6**	**86.5**

and $n_b = 3$. For VisDA-2017 dataset, we use the same optimization setting and learning rate schedule as the pretraining stage but set the initial learning rate η_0 to be 1e-4 for all convolutional layers and 1e-3 for all BatchNorm layers. We empirically set $\tau = 0.078$, $\gamma = 2$, $|\mathcal{C}'| = 6$ and $n_b = 10$. Selection of hyper-parameters will be studied in Sec. 4.3. All results reported below are the average of 3 independent runs and we manually set the random seed to guarantee reproducibility. All experiments are conducted with PyTorch and MindSpore [14] on NVIDIA 1080Ti GPUs.

4.2. Experimental results

Tabs. 1 to 3 demonstrate the experimental results of several recent SFDA methods and traditional DA methods. Best results among SFDA methods are shown in bold font. We achieve state-of-the-art performance on Office-Home (72.9%) and VisDA-2017 (86.5%). As the scale of dataset gets larger, our method performs increasingly better.

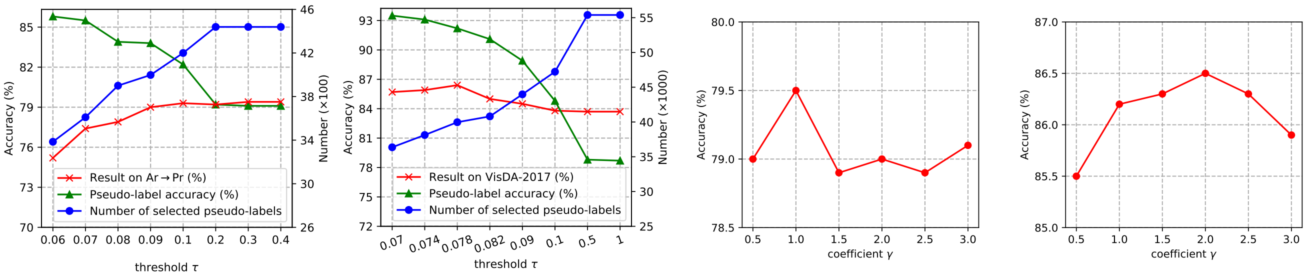
Tab. 1 shows the adaptation results on Office-31 dataset. Our method has the same best result (90.1%) as A^{2}Net [52] and is comparable to some traditional domain adaptation algorithms which require source data. Unlike Office-Home and VisDA-2017, Office-31 is a small-scale dataset whose image number of each class is around 40 on average. Therefore, it is hard for our method to accurately estimate the source distributions from statistics of target data. Yet we still achieve the best results on average and on 2 of 6 tasks.

Tab. 2 shows the results on Office-Home benchmark, in which our method achieves state-of-the-art average performance (72.9%) and performs the best on 7 of 12 transfer tasks among all the SFDA methods. Our method is even superior to some of the traditional domain adaptation methods which require source data. This dataset is larger in scale than Office-31 and thus is able to provide adequate target data to estimate source distributions more accurately.

Tab. 3 shows the per-class and average accuracy on VisDA-2017 benchmark. Our method achieves state-of-the-art performance among all SFDA methods and is higher than the second best A^{2}Net [52] by a margin of 1.1%. Despite the huge domain gap between the source domain (Synthetic) and the target (Real), our method still achieves 86.5% average accuracy due to the vast number of target images (∼55K) for estimating the source distributions, which is the key to our success. Statistics derived from sufficient of data can better reflect the real distribution.

4.3. Ablation studies

Confidence threshold τ. The precision of the estimation of class-conditioned source distributions relies on the correctness of target pseudo-labels included by $\mathcal{D}'_t$ in Eq. (2). Figs. 3a and 3b shows the sensitivity analysis on model per-



(a) Ar→Pr

(b) VisDA-2017

(c) Ar→Pr

(d) VisDA-2017

Figure 3. Analysis on hyper-parameter sensitivity. (a) and (b) Sensitivity of pseudo-labels and model performance w.r.t. τ. (c) and (d) Sensitivity of model performance w.r.t. γ.

formance, pseudo-label accuracy of $\mathcal{D}_t'$ and the number of data included by $\mathcal{D}_t'$ w.r.t. confidence threshold τ. Specifically, a small threshold τ would reject more incorrectly labeled data but the total number of data in $\mathcal{D}_t'$ would be reduced. Conversely, a large threshold will enlarge the scale of $\mathcal{D}_t'$ but introduce more false labels. Therefore, τ needs to be selected carefully. As shown in Fig. 3a, for Ar→Pr task in Office-Home dataset, despite the drop in pseudo-label accuracy caused by increasing τ, the performance of our method keeps improving in synchronization with the number of target data included by $\mathcal{D}_t'$. We conjecture that having sufficient pseudo-labeled data is more important than the accuracy of pseudo-labels to the estimation of source distributions for small-scale dataset. So we set $\tau = 0.6$ for both Office-31 and Office-Home to allow more pseudo-labels. However for VisDA-2017 dataset, as shown in Fig. 3b, the best performance is obtained when $\tau = 0.078$. Since VisDA is a large-scale dataset, a small τ can guarantee both the accuracy of pseudo-labels and the number of selected confident data simultaneously.

Covariance coefficient γ. Figs. 3c and 3d shows the experimental results with different $\gamma \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ on Ar→Pr task of Office-Home dataset and on VisDA-2017 dataset, respectively. Larger covariance matrix leads to more flexible feature activations. Thus the value of γ in Eq. (4) controls the semantic diversity of features sampled from the surrogate source distribution. By expanding the sampling range, features far from anchors can be sampled. Fig. 3d shows that the performance on VisDA-2017 is improved by a margin of 0.2% when $\gamma = 2$. However, an inappropriate value of γ may lead to a sub-optimal solution.

Estimation of the source mean. To verify the effectiveness of the estimator $\hat{\mu}_k^s$ in Eq. (3), we use several variants to replace our estimation. If we directly treat the intra-class feature mean derived from confident target data in $\mathcal{D}_t'$ as the mean of surrogate source distribution $\hat{\mu}_k^s = \bar{f}_k^t = \frac{\sum_i f_{i,k}^t}{\sum_{x_i^t \in \mathcal{D}_t'} \mathbb{1}(\hat{y}_i^t = k)}$, as shown in Tab. 4, the performance of our method decreases. Especially on VisDA-2017, the performance drops by 17.7%. On the other hand, if we directly use the anchors as estimated mean $\hat{\mu}_k^s = \mathbf{w}_k^G$, the performance becomes even worse. This suggests that information of source anchors and information of target features complement each other. In addition, *update once* in Tab. 4 means we only update both $\hat{\mu}_k^s$ and $\hat{\Sigma}_k^s$ for once at the very beginning of the whole SFDA training process, which leads to a sub-optimal result.

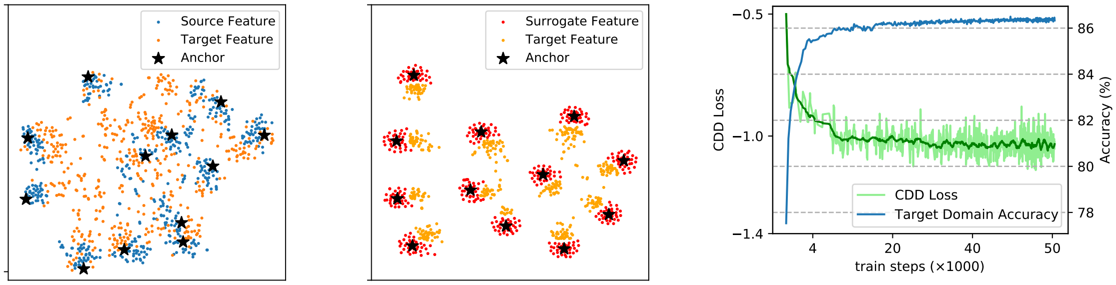
Table 4. Performance with different $\hat{\mu}_k^s$ on Ar→Cl, Ar→Pr, Ar→Rw tasks (Office-Home) and VisDA-2017 dataset.

estimator	Ar→Cl	Ar→Pr	Ar→Rw	Avg.	VisDA
$\hat{\mu}_k^s = \bar{f}_k^t$	59.2	78.0	80.2	72.5	68.8
$\hat{\mu}_k^s = \mathbf{w}_k^G$	47.1	68.8	76.2	64.0	64.1
update once	55.7	79.2	81.1	72.0	79.7
Ours	**59.7**	**79.5**	**82.4**	**73.9**	**86.5**

Table 5. Comparing with maximum probability-based pseudo-labeling method on Ar→Cl, Ar→Pr, Ar→Rw tasks (Office-Home) and VisDA-2017 dataset.

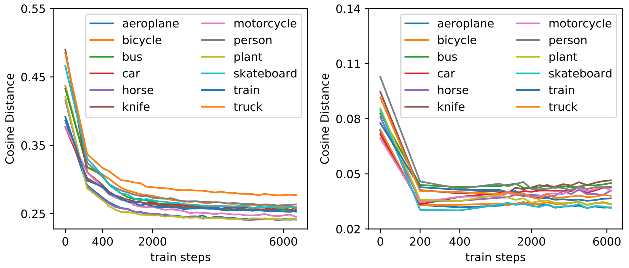
τ'	Ar→Cl	Ar→Pr	Ar→Rw	Avg.	VisDA
0.975	48.8	74.1	77.2	66.7	85.7
0.950	55.8	76.3	79.6	70.6	85.8
0.925	58.1	76.4	80.0	71.5	85.5
0.900	57.8	78.3	81.4	72.5	85.6
0.875	58.4	78.9	82.3	73.2	85.5
0.850	59.0	78.8	81.6	73.1	85.3
Ours	**59.7**	**79.5**	**82.4**	**73.9**	**86.5**

Robustness of pseudo-labeling strategy. Obtaining robust pseudo-labels is important to the following SDE process, since high-quality pseudo-labels can provide accurate estimation for the mean and covariance of each distribution. If pseudo-labels are corrupted, the estimated distribution would be diverged from the real distribution, which makes the sampled surrogate features unable to represent the real source features of a certain class. To validate the robustness of our anchor-based spherical k-means clustering pseudo-labeling method, we conduct experiments and show the results in Tab. 5. Instead, we use a maximum probability-based strategy to assign pseudo-labels: $\hat{y}_i^t = \arg\max_k \sigma_k(\mathbf{G}(\mathbf{F}(x_i^t)))$, where σ is the K-way softmax function that generate probabilities for each class. We also set a threshold τ' to select confident samples



(a) Before Adaptation (b) After Adaptation (c) Training Curves

Figure 4. Visualization on VisDA-2017 dataset. (a) T-SNE visualization of source features and target features before SFDA. (b) T-SNE visualization of surrogate features and target features after SFDA. (c) Curves of CDD loss and model performance on target domain.



(a) (b)

Figure 5. (a) Cosine distance between the centers of target features $\bar{f}_k^t$ and source anchors $\mathbf{w}_k^G$ for each class. (b) Cosine distance between target covariance Σ_k^t and corresponding source covariance Σ_k^s for each class.

whose maximum probabilities are greater than τ' to construct the confident target dataset $\mathcal{D}_t'$. Multiple values of τ' are tested to guarantee a fair comparison. Tab. 5 shows that our anchor-based clustering pseudo-labeling method outperforms maximum probability-based method on both Office-Home dataset and VisDA-2017 dataset.

4.4. Visualization and empirical analysis

We visualize the experimental results on VisDA-2017 dataset and analyse the proposed SFDA-DE method.

Training curves. Fig. 4c shows the training curves of CDD loss and model accuracy on target domain during source-free adaptation process. Our method converges stably and shows superior performance from an early stage.

Domain shift. We utilize t-SNE visualization to demonstrate the distributions of feature representations in both source and target domains. As shown in Fig. 4a, a large amount of target data (represented by orange dots) disperses in the feature space before adaptation due to severe domain shift problem while source data (represented by blue dots) gathers around the anchors and forms intra-class clusters.

Visualization of surrogate features. Red dots in Fig. 4b represent the surrogate features derived from SDE with covariance multiplier $\gamma = 2$, which enlarges the sampling range. These surrogates are distributed around corresponding anchors to simulate source features of each class.

Effectiveness of our method. After SFDA training, as shown in Fig. 4b, target features are pulled towards corresponding anchors and merged into the surrogate feature clusters. Besides, low density area can be clearly observed in the feature space after adaptation. This suggests our SFDA-DE method can learn discriminative features for unlabeled target domain without using source data.

Calibration of distribution mean. In SDE, anchors are utilized to calibrate the mean of estimated source distribution according to Eq. (4), since target features drift away from source features at the early stage of training. Therefore, target class centers $\bar{f}_k^t = \frac{\sum_i f_{i,k}^t}{\sum_{x_i^t \in \mathcal{D}_t'} \mathbb{1}(\hat{y}_i^t = k)}$ cannot serve as a good estimator of μ_k^s. As shown in Fig. 5a, the distance between target feature centers and source anchors is diminished as training proceeds. Target features gradually approach the corresponding anchors of the same class, which means the calibration of $\hat{\mu}_k^s$ is effective.

Estimation bias of covariance. Fig. 5b visualizes the classwise estimation bias of distribution covariance $\hat{\Sigma}_k^s = \Sigma_k^t$ over the ground truth source covariance Σ_k^s w.r.t. training steps. The gap in between is mitigated in the early stage of training and is kept at a low level, which verifies our assumption made in Sec. 3.3. Thus, class-conditioned source covariance can be approximated via high-quality pseudo-labeled target data.

5. Conclusions

In this paper, we propose a novel framework named SFDA-DE to address source-free domain adaptation problem via estimating feature distributions of source domain in the absence of source data. We utilize domain knowledge preserved by source anchors to obtain high-quality pseudo-labels for target data to achieve our goal. Sufficient experiments validate the effectiveness and superiority of our method against other strong SFDA baselines.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant No. 61876007.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007. 1, 2

[2] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[4] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2020. 1, 5, 6

[5] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, Xilin Chen, and Xuelong Li. Flowing on riemannian manifold: Domain adaptation by shifting covariance. *IEEE transactions on cybernetics*, 44(12):2264–2273, 2014. 4

[6] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, number 6, 2018. 6

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 2

[10] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6

[11] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Classification-aware semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 964–965, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[13] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4043–4052, 2020. 5, 6

[14] Huawei. Mindspore. https://www.mindspore.cn/, 2020. 6

[15] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538. PMLR, 2014. 2

[16] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020. 5, 6

[17] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 2, 4, 5

[18] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2021. 2, 5, 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1, 2

[20] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 2

[21] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 6

[22] Limin Li and Zhenyue Zhang. Semi-supervised domain adaptation by covariance matching. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2724–2739, 2018. 4

[23] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 2, 5, 6

[24] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11516–11525, 2021. 1, 4, 6

[25] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2, 3, 4, 5, 6

[26] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2

[27] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound

domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[29] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[30] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in neural information processing systems*, pages 1640–1650, 2018. 1, 2

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in Neural Information Processing Systems*, 29:136–144, 2016. 2, 4

[32] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2, 4

[33] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 6

[34] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1094–1103, June 2021. 1, 6

[35] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 2

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 2

[37] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 2, 5

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2, 5

[39] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 2

[40] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[42] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009. 2

[43] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pages 1433–1440. Citeseer, 2007. 1, 2

[44] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008. 2

[45] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 2017. 2

[46] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 2

[47] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020. 5, 6

[48] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *arXiv preprint arXiv:2103.14357*, 2021. 2

[49] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2, 5

[51] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32:12635–12644, 2019. 4

[52] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021. 2, 5, 6

[53] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2018. 2

[54] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435, 2019. 6

[55] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020. 6

[56] Yixing Xu, Kai Han, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Learning frequency domain approximation for binary neural networks. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[57] Yixing Xu, Yunhe Wang, Hanting Chen, Kai Han, Chunjing Xu, Dacheng Tao, and Chang Xu. Positive-unlabeled compression on the cloud. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[58] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 2, 4, 6

[59] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021. 2

[60] Werner Zellinger, Bernhard A Moser, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Robust unsupervised domain adaptation for neural networks via moment alignment. *Information Sciences*, 483:174–191, 2019. 2

[61] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 5